IDIAP RESEARCH REPORT

FEATURE MAPPING USING FAR-FIELD MICROPHONES FOR DISTANT SPEECH RECOGNITION

Ivan Himawan Petr Motlicek David Imseng
Sridha Sridharan[a]

Idiap-RR-20-2016

AUGUST 2016

[a]QUT

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Feature Mapping using Far-Field Microphones for Distant Speech Recognition

Ivan Himawan[a], Petr Motlicek[a], David Imseng[a], Sridha Sridharan[b]

[a]*Idiap Research Institute, Martigny, Switzerland*
[b]*Queensland University of Technology, Australia*

Abstract

Acoustic modeling based on deep architectures has recently gained remarkable success, with substantial improvement of speech recognition accuracy in several automatic speech recognition (ASR) tasks. For distant speech recognition, the multi-channel deep neural network based approaches rely on the powerful modeling capability of deep neural network (DNN) to learn suitable representation of distant speech directly from its multi-channel source. In this model-based combination of multiple microphones, features from each channel are concatenated and used together as an input to DNN. This allows integrating the multi-channel audio for acoustic modeling without any pre-processing steps. Despite powerful modeling capabilities of DNN, an environmental mismatch due to noise and reverberation may result in severe performance degradation when features are simply fed to a DNN without a feature enhancement step. In this paper, we introduce the nonlinear bottleneck feature mapping approach using DNN, to transform the noisy and reverberant features to its clean version. The bottleneck features trained on clean signal are used as a teacher signal because they contain relevant information to phoneme classification, and the mapping is performed with the objective of suppressing noise and reverberation. The individual and combined impacts of beamforming and speaker adaptation techniques along with the feature mapping are examined for distant large vocabulary speech recognition, using a single and multiple far-field microphones. As an alternative to beamforming, experiments with concatenating multiple channel features are conducted. The experimental results on the AMI meeting corpus show that the feature mapping, used in combination with beamforming and speaker adaptation yields a distant speech recognition performance below 50% word error rate (WER), using DNN for acoustic modeling.

Keywords: deep neural network, bottleneck features, distant speech recognition, meetings, AMI corpus

1. Introduction

Automatic speech recognition from distant microphones is a challenging task, because the speech signals to be recognized are degraded by the presence of interfering signals and reverberation due to large speaker-to-microphone distance [1]. The conventional multi-channel enhancement techniques, such as beamforming, are widely employed to suppress noise and reverberation from the desired speech when multiple microphones (e.g., microphone arrays) are used to capture audio signals [2, 3].

In the context of ASR, the conventional speech enhancement methods are typically used as a pre-processing step to reduce mismatch between a model trained using clean speech and the noisy features. Since these methods are designed to improve signal-to-noise ratio (SNR), or signal-to-interference-plus noise ratio, the performance of the speech recognizer will be sub-optimal. In case of multi-channel ASR, there have been studies on designing a beamformer with the aim of optimizing ASR performance. A technique such as likelihood maximizing beamforming (LIMABEAM) [4, 5] specifically optimizes array parameters using gradient descent to maximize the likelihood of the recognized hypothesis under an ASR speech model, given the filtered acoustic data. Recent research on LIMABEAM suggests no significant improvement using the standard LIMABEAM on large vocabulary distant speech recognition on the AMI meeting corpus and it is recommended to use a better optimization strategy for any LIMABEAM implementation [6].

Further, it is also possible to perform recognition from microphone arrays without employing any pre-processing steps. For example, each individual chan-

Email addresses: ihimawan@idiap.ch (Ivan Himawan), pmotlic@idiap.ch (Petr Motlicek), dimseng@idiap.ch (David Imseng), s.sridharan@qut.edu.au (Sridha Sridharan)

nel can be separately recognized, and the recognition hypotheses are combined using a confusion network combination to select a word sequence with the highest probability [7, 8]. Channel selection approaches such as finding the channel producing the maximum acoustic likelihood [9], or selecting the channel with the maximum confidence from its decoded sequence [10], may be particularly useful when microphones are loosely specified in users' environments. Since recognition needs to be performed before any hypothesis is selected or combined, these decoder-based approaches for recognizing multiple microphones are computationally demanding (i.e., multi-pass-systems) .

Recently, acoustic models based on DNN have been shown to significantly improve the ASR performance on a variety of tasks when compared to the conventional Gaussian mixture model hidden Markov model (GMM/HMM) systems. Several international challenges have recently been organized to attract researchers' interest in providing the ASR solution in reverberant environments, such as the ASpIRE [11] and CHiME challenge series [12, 13]. In those challenges, participants were encouraged to build state-of-the-art speech recognition systems that are robust to various environmental factors and recording scenarios, while minimizing the impact of mismatch between training and testing conditions. For example, the recent 3rd CHiME challenge specifically addressed the far-field recordings from a mobile tablet device, captured using six microphones positioned around the tablet frame in real-world environments. It was reported that one of the most effective techniques, where significant gains have been achieved, is to transform the DNN features using feature-space maximum likelihood linear regression (fMLLR) [14, 15], and some of the best scoring systems have used baseline DNN configurations for acoustic modeling [13]. Apart from DNN's superior modeling capacity in acoustic modeling, a DNN which is trained with context-dependent phonetic targets can be used to produce neural-network-based features or bottleneck (BN) features. These features have been shown to be effective in improving the performance of ASR systems especially when exploited in combination with traditional short-term spectral features, such as MFCCs or PLPs [16, 17]. The BN features are usually extracted from one of the internal layers of DNN (with a small number of hidden units in comparison to the size of the other layers) and represent a nonlinear transformation (while usually reducing dimensionality) of the input features [18, 16]. The stacked BN features which are extracted from the cascaded DNN structures have been investigated on several ASR tasks, such as speech recognition of Cantonese spontaneous telephone conversations [19] and speech recognition with minimum resource [20]. In [21], the BN features were also used for far-field speech recognition.

This paper introduces a nonlinear BN feature mapping approach by using the BN feature of a close-talking microphone (referred to as the individual headset microphone (IHM)) as a target for distant speech input. The DNN is used to map the noisy and reverberant features to the BN-based features extracted from the close-talking input. Once the mapping is completed, the transformed BN features are extracted for training a new acoustic model [22]. The model-based combination of multiple microphones using the transformed BN features is proposed to integrate the multi-channel inputs for acoustic modeling. For the feature mapping approach, the fMLLR for speaker adaptation is applied to the features prior to DNN training and to the transformed BN features in the stacked hybrid fashion [23]. The fMLLR has been shown to be effective in both hybrid and tandem DNN-based systems for removing speaker variabilities and variations in the recording process, due to speaker-to-microphone distances and the use of different microphone channels [24, 23, 25]. Although many recent speaker adaptation techniques for DNN have been proposed such as learning hidden unit contributions (LHUC) [25], providing speaker identity vectors (i-vectors) along with regular ASR features as input to neural nets [26, 27], and incorporating i-vectors to project the speech features into a speaker-normalized space [28, 29], it is straightforward to use fMLLR in DNN/HMM hybrid acoustic models. The GMM/HMM models which are usually trained to generate the alignment with context-dependent phone states for DNN training can further be used to estimate speaker transforms. This paper investigates the feature mapping approach for far-field microphones by examining the individual and preferably combined impacts of beamforming and fMLLR for robust ASR. The comparison to multi-condition training is also presented.

This paper is organized as follows. Section 2 discusses related work. Section 3 describes the DNN-based mapping approach. The experimental setup is described in Section 4. The ASR results, employing the BN feature mapping approach using far-field microphones, are presented in Section 5. Section 6 discusses the results. Finally, the study is concluded in Section 7.

2. Related Work

2.1. Speech Enhancement using DNN

In a noisy and reverberant room, the reverberated speech $x(t)$ is represented in time domain as the convolution of the clean speech signal $s(t)$ and the room impulse response $h(t)$, corrupted by additive noise $n(t)$, as

$$x(t) = s(t) * h(t) + n(t). \quad (1)$$

The effect of early reflection and late reverberation on the reverberant signal is considered as a separate process in many studies. The late reverberation part of the room impulse response is often modeled as an exponentially damped Gaussian noise process and treated as additive noise. Hence, the observed reverberant signal $x(t)$ can be written by using the notation in [1] as

$$x(t) = s(t) * h_e(t) + r(t) + n(t), \quad (2)$$

where $h_e(t)$ is the early reflection part of the impulse response and $r(t)$ is the late reverberation component of $x(t)$.

The conventional methods to recognize reverberated speech captured from distant microphones is to first reconstruct a clean version of the speech. This may be performed with a blind dereverberation method, such as estimating the inverse filter solely on the observed signals capable to cancel out the reverberation effects [30, 31]. Since the late reverberation is often treated as additive noise, speech enhancement methods, such as spectral subtraction [32] and minimum mean-square error (MMSE)-based techniques [33, 34], may be used to mitigate the impact of reverberation. If two or more microphones are used to capture speech, multi-channel speech enhancement techniques such as multi-channel Wiener filter [35], beamforming followed by post-filtering [36], or blind speech separation [37] can be used for improving the quality of speech. One drawback of these conventional speech enhancement methods is that they often fail to track the non-stationary noise signals in real-world scenarios.

One of the emerging speech enhancement approaches is based on deep architectures. In [38], the DNN-based regression model was trained using noisy data and their corresponding clean speech version. The developed model was then used to predict the clean speech features. Improvements were reported across different noise conditions where the DNN-based speech enhancement was shown to be effective for dealing with non-stationary noises in real-world environments. The speech enhancement may be formulated as a binary classification problem to estimate the ideal binary mask (IBM), which is used to attenuate the energy within the noise dominant time-frequency units. For robust ASR, the ideal ratio mask (IRM), defined as the ratio of speech energy to total energy (speech and noise) in each time-frequency unit, has been shown to have a better performance compared to using IBM in a large vocabulary speech recognition task [39]. In [40], the DNN is used to estimate the instantaneous SNR for computing IRM, subsequently applied to filter out noise from a noisy Mel spectrogram. The recurrent neural networks (RNNs), with their ability to model the temporal dependencies in speech, have also been employed to estimate the time-frequency masks from the magnitude spectrum of a noisy signal for speech enhancement and recognition [41]. For speech recognition applications, the speech enhancement approaches are typically exploited as front- end processing to reconstruct the clean version of the speech, which is then fed into a speech recognizer.

2.2. Multi-channel integration in acoustic modeling

In speech processing and especially ASR applications, the use of microphone arrays instead of close-talking microphones is popular, since they enable natural interactions between the speaker and devices. The multi-channel speech-enhancement-based approaches can be employed for this purpose, where the task of enhancement and recognition is performed separately in a cascaded fashion, instead of being integrated into one system. A basic technique such as delay-sum beamforming works by compensating delays from the individual microphone channels so that the target signal from a particular direction synchronizes, while noises are canceled through destructive interference. The relative time delays between channels are typically estimated using generalized cross correlation with phase transform (GCC-PHAT) [42] and processed using Viterbi post-processing to select the reliable delays, while minimizing undesired beam-steering toward interfering events [3]. The beamformed audio may then be enhanced using post-filtering [36] before it is passed to a speech recognizer as single-channel speech. Other sophisticated beamforming methods, such as generalized sidelobe canceller, perform spatial filtering while, at the same time, reduce the influence of noise at the location of interest. These advanced techniques take into account the estimated noise or interfering signal characteristics for superior noise suppression capability [43, 44]. In the context of ASR, beamforming techniques have been successfully exploited in the ICSI/SRI [45] and AMIDA [46] systems for transcriptions of meetings [47].

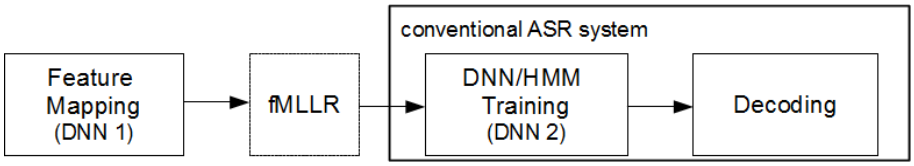

Figure 1: Block diagram of the DNN-based feature mapping approach. The overall approach comprises two DNNs, first employed in feature mapping (DNN 1) and the second as an acoustic model in conventional ASR system (DNN 2).

Another research efforts have explored unified multi-channel-based speech recognition such as LIMABEAM and multi-channel-based neural networks speech recognizer. In the LIMABEAM approach, a filter-and-sum beamforming structure is employed in which the parameters of the beamforming filter are optimized using the gradient descent technique so that the filtered signal will generate a sequence of features that maximize the likelihood of correct transcriptions. On the other hand, the multi-channel DNN-based speech recognizer performs a direct concatenation of multi-channel features (using standard PLP features [48] or in combination with BN features [21]). On the AMI corpus, channel concatenation is shown to perform better than applying beamforming for a small number of microphones. This indicates that the DNN is able to learn representation of distant speech directly by using multi-channel input [49]. The channel-wise convolution followed by a cross-channel max pooling using convolutional neural network (CNN) is proposed in [50] for selecting the best features within the channels. It was shown that a CNN with the proposed configuration trained directly on the output of multiple microphones yields higher speech recognition accuracy when compared to a CNN trained on the output of a delay-sum beamformer. Since the multi-channel features are directly used for acoustic modeling, these multi-channel integration approaches are not specifically designed to suppress noise and reverberation.

A recent study in advanced acoustic modeling using deep long short-term memory (LSTM) recurrent neural networks reported significant improvement for AMI's single distant microphone (SDM) task with 47.7% WER, even though it does not consider multi-channel inputs [51]. This work may not be directly compared with our results since it used sequence discriminative training with dropout and DNN to force align the training data to generate labels for LSTM training.

2.3. Feature mapping techniques using DNN

Nonlinear modeling techniques such as neural networks can be employed to model the complex noise

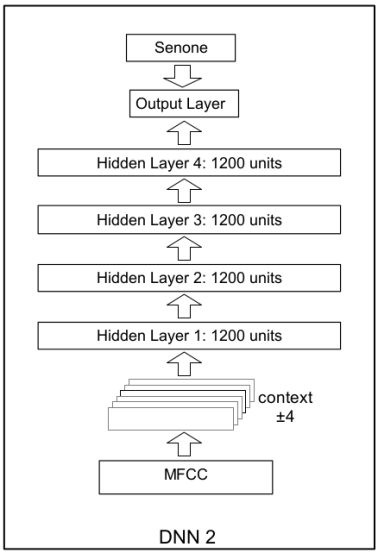

Figure 2: DNN 2 architecture with four hidden layers.

corruption process by learning the mapping function between noisy speech and its clean version. The mapping is performed between features extracted from noisy and clean speech signals to obtain an optimal set of parameters through the error backpropagation algorithm [52, 53, 54]. The goal is to obtain clean or enhanced speech from the noisy input via a nonlinear transformation using neural networks such as a deep denoising autoencoder [55] or a multilayer perceptron (MLP) [56]. A recent study [56] considers a multi-speaker scenario where the nonlinear mapping is performed with the objective of improving overlapped speech recognition using beamformed audio from a microphone array. The mapping is performed in the log mel-filterbank energy domain by minimizing the minimum mean squared error as an objective function. In order to improve the quality of the estimated clean speech, the feature mapping can be performed from multiple beamformed sources. For example, two beamformers with binary masking post-filters directed to the target audio signals and to the interfering speech, respectively, such that the interfering components in the features ex-

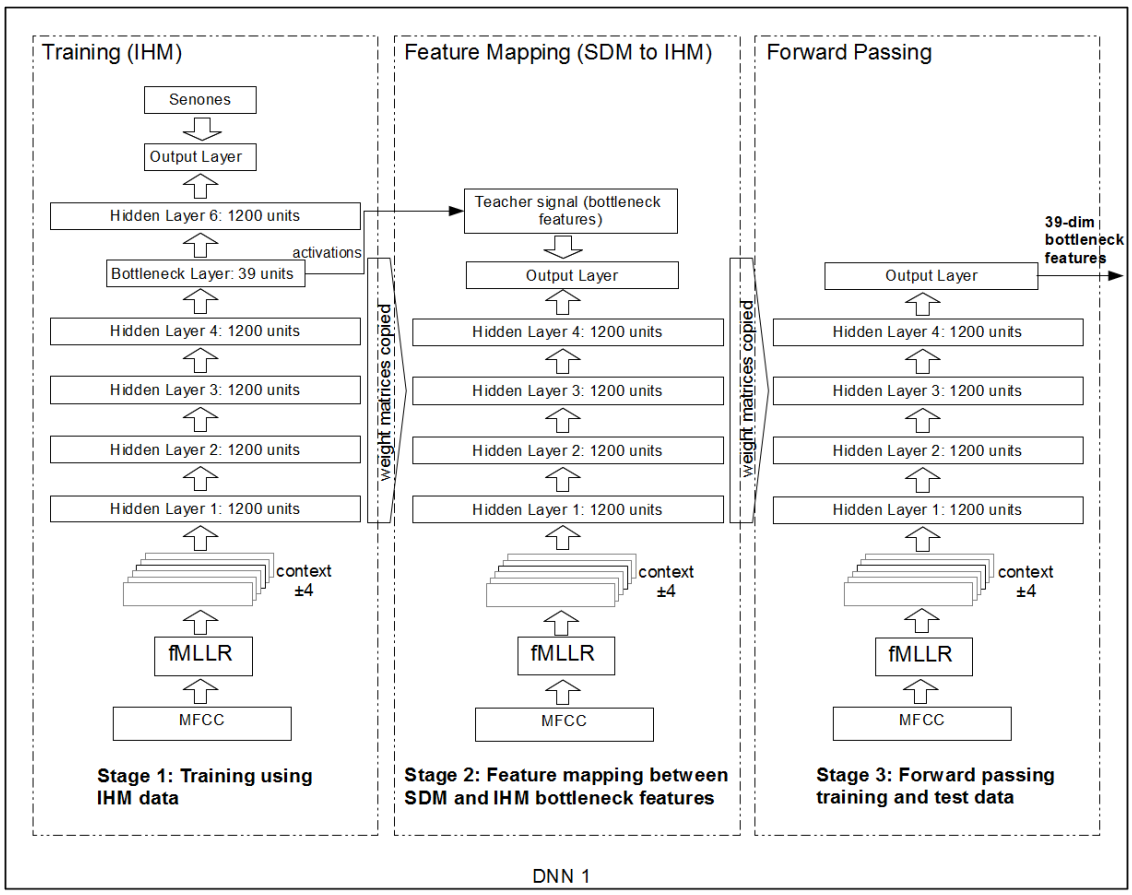

Figure 3: Feature mapping with fMLLR for single distant microphone. The fMLLR transform is applied on the IHM training data prior to training the DNN 1.

tracted from the desired target signals can be subtracted by features extracted from interfering speech. While significant improvements in speech recognition accuracy are reported if the new acoustic model is trained from the estimated clean features, the experiments are limited to the recognition of digits.

The novelty of our framework in comparison to other neural net-based mapping is the use of BN features as the teacher signal, which is obtained from the BN layer of a DNN. A DNN is trained to estimate the posterior probability for a set of ASR states called *senones* (clustered, context-dependent sub-phonetic HMM states generated by a set of phonetic decision trees) given acoustic input. In contrast to other feature mapping approaches, our method does not require reconstruction of the estimated clean speech features prior to training a new acoustic model, which are rather designed for speech enhancement. This paper considers a model-based combination of multiple microphones. Our previous work of single channel mapping in [22] is extended, and results are compared with conventional speech enhancement techniques for distant large vocabulary speech recognition. Further, we investigate fMLLR transform for speaker adaptation when it is used within the feature mapping framework, and show that the feature mapping is complementary to fMLLR feature space adaptation. Finally, this work is related to multi-condition training in which a DNN is trained with speech signals in multiple conditions, in order to deal with different acoustic channels and various environmental noises [17, 57]. We hypothesize that features extracted from the hidden layers from such networks are inherently robust to noise. A comparison between the mapping approach and DNN trained using BN features extracted from the multi-condition network is presented in this paper.

3. DNN-based Feature Mapping

The block diagram of the proposed DNN-based feature mapping is shown in Figure 1. It consists of feature mapping, the fMLLR transform estimation, which can be applied on the transformed features, and a DNN/HMM ASR system. The DNN/HMM hybrid configuration used in this paper is based on a DNN trained to estimate the emission probabilities of the HMM states. In the baseline experiments, the DNN comprises four 1200-neuron hidden layers. Figure 2 shows a DNN architecture used in this study in more detail.

For the feature mapping approach, the input features are extracted from the single channel output (either from a distant microphone or after microphone array beamforming). Input features at frame n are denoted as vectors $\mathbf{x}(n)$. The aim is to find the estimate of BN features of clean speech, $\hat{\mathbf{c}}(n)$, using a multilayer perceptron with multiple hidden layers with a set of parameters, $\theta = \{\mathbf{W}_1, \ldots, \mathbf{W}_L, \mathbf{b}_1, \ldots, \mathbf{b}_L\}$, which denote all optimal weights and bias parameters. We use the notation from [58], $\hat{\mathbf{c}}(n)$ is obtained as

$$\begin{aligned} \mathbf{u}_l &= \sigma(\mathbf{W}_l \mathbf{u}_{l-1} + \mathbf{b}_l), \quad \text{for } 1 \leq l < L \\ \hat{\mathbf{c}}(n) &= \mathbf{W}_L \mathbf{u}_{L-1} + \mathbf{b}_L, \end{aligned} \tag{3}$$

where $\mathbf{u}_l$ is the input to the $l+1$-th layer, $\mathbf{W}_l$ denotes the matrix of connection weights between $l-1$-th and l-th layers, $\mathbf{b}_l$ is the additive bias vector at the l-th layer, and $\sigma(.)$ is the sigmoidal activation function. The θ is obtained by minimizing the mean squared error objective function

$$E = \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{c}(k) - \hat{\mathbf{c}}(k)\|^2 \tag{4}$$

over K training examples (i.e., frames), where $\mathbf{c}(k)$ denotes vectors of teacher BN features generated from clean speech.

A DNN/HMM hybrid network, which is originally trained to minimize a negative log posterior probability cost function over the set of training examples, is used to provide the optimal parameters θ using Equation 4. The final BN feature estimates of clean speech $\hat{\mathbf{c}}(n)$ can be obtained by forward-passing the reverberant feature vectors through the network with optimized parameters θ.

Compared to our previous work in [22], the first fMLLR transform is applied to the input features prior to training the first DNN. Hence, we use speaker-normalized distant-talking speech features as input for the mapping procedure. In addition, the second fMLLR transform is also applied to the BN features after feature mapping. The feature mapping approach consist of three stages, as shown in Figure 3 and described below:

- **First stage: training using IHM data**

  Using IHM data, the DNN is trained on the fMLLR adapted features to generate BN features. For the experiments in this paper, we apply a DNN with eight layers (i.e., six hidden layers including the BN layer). The BN layer is placed just between the 4^{th} and the 6^{th} hidden layer and has 39 dimensions with linear activation functions.

- **Second stage: feature mapping between SDM features and IHM BN features**

  The BN features extracted from the previous DNN are used as a teacher input for a nonlinear feature transformation of distant-talking speech input. In other words, the BN features trained on IHM data are used to transform the parameters of the reverberated speech to a new space close to clean speech. To learn this mapping, the network is trained using the standard error backpropagation procedure and the optimization is done through stochastic gradient descent by minimizing the mean squared error objective function using Equation 4.

- **Third stage: forward passing training and test data**

  Once the mapping is learned, the trained network structure is used to generate new speech features from the activations of the units of the output layer for training the new acoustic model. This yields 39-dimensional transformed features (to compare with 39 MFCCs). The fMLLR transform is applied on these new features before training the second DNN. This DNN has four 1200-neuron hidden layers, as shown in Figure 2. For decoding, test sets for IHM and SDM are fed to the trained network, and the transformed features are extracted from the activations of the output layer.

4. Experimental Data and Setup

The ASR experiments employ both the headset and their corresponding distant microphone recordings from the AMI meeting corpus[1], which contain meetings recorded in instrument-equipped meeting rooms at three

[1] http://groups.inf.ed.ac.uk/ami/download

sites in Europe (Edinburgh, IDIAP, and TNO). The single distant microphone and the multiple distant microphones (MDM) of the primary array were used. The SDM was represented by the first microphone and the MDM used either 4 or 8-channel inputs from the primary microphone array. There are about 67 hours of training data and around 7 hours of evaluation data available (after performing voice activity detection). In addition to the AMI test set, the trained acoustic models are evaluated on a NIST Rich Transcription (RT-07) ASR evaluation task to determine if the feature mapping approach trained on the AMI corpus improves the ASR performance of unseen condition. The experiments used the suggested AMI corpus partitions for training and evaluation sets [46, 47], even though some of the meeting recordings were discarded from the original corpus when array recordings were missing, to ensure that both headset recordings and the corresponding synchronized array recordings are available for training and testing.

Our previous work showed that SDM system trained using alignment generated from IHM (clean) ASR system provided significantly better performance [22], compared to SDM system trained using alignment from SDM. Since SDM data are synchronized with IHM data (on a frame-level), the SDM models are trained using HMM state alignments generated for IHM recordings.

The Kaldi toolkit is used for training DNN/HMM systems and for generating fMLLR features using the provided training scripts [59]. The IHM and SDM ASR configurations are trained on 39-dimensional MFCC features, including their delta and acceleration versions. The DNNs for both configurations are trained to estimate posterior probabilities of roughly 4K tied-state (senone) targets. The DNNs use a 9-frame temporal context, enriched with cepstral mean only and cepstral mean and variance normalization for fMLLR and non-fMLLR systems, respectively. The AMI pronunciation dictionary, of approximately 23K words, is used in the experiments and the Viterbi decoding is performed using a 2-gram language model (LM) [60], previously built for NIST RT-07 corpora [46]. An additional experiment with a stronger LM (4-gram) is performed with the best system to determine if the gains in acoustic modeling are retained.

		Test sets		
	Trained on	IHM	SDM	RT-07
Single-condition SDM	IHM	32.3	*76.0*	37.2
	+ fMLLR	28.4	*74.7*	33.4
	SDM	*46.7*	58.0	50.8
	+ fMLLR	*43.2*	54.2	48.0
	Bottleneck-based systems:			
	IHM	33.8	*63.4*	38.8
	+fMLLR	31.2	*57.0*	36.7
	SDM	*41.8*	57.3	46.7
	+fMLLR	*40.3*	52.5	45.7
Single-condition MDM		MDM		
	4*bmit*	55.1		
	4*fmcct*	55.3		
	4*fmbmit*	54.0		
	8*bmit*	52.7		
	8*fmcct*	54.8		
	8*fmbmit*	51.8		
	+fMLLR	49.9		
	+fMLLR (4gram-LM)	47.4		
Multi-condition		IHM	SDM	
	Multi-condition	33.3	59.5	
	+fMLLR	29.5	56.2	
	Bottleneck-based systems:			
	IHM	31.3	*65.5*	
	+fMLLR	26.9	*55.4*	
	SDM	*36.8*	56.8	
	+fMLLR	*30.3*	52.2	

Table 1: WERs[%]: Baseline and mapping results with single and multiple distant microphones for single and multi-condition systems. In order to simplify comparison, the ASR results for mismatch conditions are displayed in *italics*.

5. Experimental Results

5.1. Single-condition Baseline

As shown on the top part of Table 1, the performance gap between IHM and SDM is large (about 44% WER absolute) for the model trained on IHM, due to the difficulty of recognizing the distant microphone. We introduced the feature mapping technique used in combination with fMLLR to reduce the mismatch between clean and reverberant conditions. Apart from improving ASR performance due to the mismatch of conditions, the feature mapping approach is also investigated for improving distant speech recognition performance (i.e., using the best model to recognize the SDM test). In order to simplify comparison, the ASR results for mismatch conditions are displayed in *italics*.

Employing fMLLR to the input features prior to

training DNN improves the ASR performance on both matched and mismatched conditions. The performance improves for the IHM model by 3.9% and 1.3% absolute WER when recognizing IHM and SDM tests, respectively. The SDM model improves the performance by 3.5% and 3.8% absolute WER when recognizing IHM and SDM tests, respectively. For RT-07 evaluation task, the performance improves by 3.8% and 2.8% absolute WER for model trained using IHM and SDM, respectively. For single-condition baseline, the best system for recognizing the IHM test by using a stronger LM (4-gram) yields 25.6% WER instead of 28.4%.

5.2. Single-condition Mapping using SDM

The results for systems without applying fMLLR have been previously reported in [22]. Compared to the baseline performance, BN-based system improves the performance on SDM while trained on IHM data by 12.6% absolute WER (from 76.0% to 63.4%; 16.5% relative), whilst a minor degradation of 1.5% absolute (4.5% relative) is observed on the matched condition. Using the model trained on SDM data yields improvement by 4.9% absolute WER (10.5% relative; from 46.7% to 41.8%) when recognizing the IHM test. This suggests that the SDM features have more discriminant classification ability, close to the IHM condition, after being transformed by the mapping network. A minor improvement of 0.7% (from 58.0% to 57.3%) absolute WER is observed when recognizing the SDM test. A similar trend is also observed when recognizing RT-07 evaluation task using the SDM model, which yields 4.1% absolute WER (8% relative; from 50.8% to 46.7%) improvement, while using the IHM model degrades the ASR performance by 1.6% absolute (from 37.2% to 38.8%).

The use of speaker adaptation improves the overall performance of the feature mapping approach. Table 1 shows that, on matched condition, using the IHM model with fMLLR gains 2.6% absolute WER (from 33.8% to 31.2%) over the non-fMLLR system, while using the SDM model gains 4.8% absolute WER (from 57.3% to 52.5%). For the BN-based system trained on SDM, a noticeable improvement of 1.7% (from 54.2% to 52.5%) absolute WER is observed when recognizing the SDM test. An improvement by 2.3% (from 48.0% to 45.7%) absolute WER is observed when recognizing RT-07 evaluation task. We attributed these improvements to the use of SDM with speaker-normalized features as input to learn the mapping, where the clean BN features act as the teacher signal. Note that the mapping results denoted as "+fMLLR" stand for ASR systems applying two-stage fMLLR transforms (first before feature mapping, second on the transformed features).

5.3. Single-condition Mapping using MDM

Beamforming has been a popular speech enhancement technique for distant speech recognition tasks. This paper further investigates the feature mapping approach from multiple distant microphones by comparing conventional beamforming, feature mapping of the beamforming signal, and feature concatenation used in combination with feature mapping. For the MDM experiments, the BeamformIt toolkit [3] (with default settings) is used to perform noise cancellation with delay-sum beamforming.

The four and eight SDM channels are also used together to perform an adaptation. More precisely, the same BN DNN (DNN 1) already trained is used to forward-pass the SDM channels (channels 1, 2, 3, and 4 from a primary array for 4-channel adaptation, and all channels from a primary array for 8-channel adaptation).

Since the input dimension of DNN is fixed, stacked multi-channel inputs would not generalize to setups with a different number of channels. Therefore, we decided to train the first DNN with 39 dimensional MFCCs for IHM data rather than with higher number of dimension for the multi-channel mapping. The 1-channel feature mapping DNN enables us to easily transform features for each microphone of the array for any number of channels. The transformed features from the four SDM channels are then concatenated to construct a new feature vector to train final DNN/HMM. Note that, for feature concatenation, no beamforming signal processing is involved. In addition, we performed feature mapping directly on the single channel of the enhanced signal. For these experiments, fMLLR is applied only to the best performing system and evaluated by using a 4-gram LM, with eight microphones.

The middle part of Table 1 (single-condition MDM) shows results of feature mapping used in combination with channel concatenation (denoted as "*fmcct*"). The results of feature mapping used with beamformed output (denoted as "*fmbmit*") are presented in the same table. Channel concatenation generates large feature dimensionality and requires a large amount of parameters for acoustic modeling. For this reason, the HMM state alignment for training the acoustic model is obtained from IHM. To have a fair comparison, all experiments using MDM exploit the IHM system to generate HMM state alignment.

5.4. Multi-condition Baseline

One of the widely-used models for noise-robust ASR is obtained from multi-style training, where examples of clean and noisy speech, under various conditions, are included in the training data. Experimental studies reported that ASR performance from multi-condition models is better in various SNR conditions when compared to a model trained only on clean speech data [61, 57]. In this paper, the multi-condition model is obtained by training IHM and SDM data together using DNN 2.

Compared to the single-condition model for recognizing matched condition, small performance degradation is observed for the multi-condition model. The bottom part of Table 1 (multi-condition) shows that the performance degrades by about 1% (32.3% compared to 33.3%) absolute WER, when evaluated on an IHM test set, and by 1.5% (58% compared to 59.5%) when evaluated on an SDM test set. A similar trend is also observed when fMLLR is applied to the input features prior to multi-style training. The small degradation in performance suggests the inherent robustness of DNN to noise. Therefore, we conducted additional experiments, described in the next subsection, where BN features extracted from a multi-condition network were used to construct the deep BN features-based DNN.

5.5. Multi-condition BN-based System

If the multilayered networks are regarded as a cascaded sequence of feature extractors followed by a logistic regression classifier at the output layer [58], it is reasonable to assume that the features extracted from the hidden layers contain information for classification and environmental noise level. Training and decoding on features extracted from such a network are assumed to be inherently robust to noise.

For the multi-condition BN-based system, instead of training feature mapping DNN, the multi-condition DNN is trained with a BN layer. Once this network with a BN layer is trained, we forward-pass the IHM or SDM training data in order to extract 39-dimensional features from the activation of the units of the BN layer. The bottom part of Table 1 shows the results from training the IHM and SDM models from such BN features using DNN 2. Compared to multi-condition with fMLLR on matched condition, the BN-based systems with fMLLR yield improvement by 2.6% (from 29.5% to 26.9%) and by 4% (from 56.2% to 52.2%) absolute WER when recognizing the IHM and SDM test sets, respectively.

6. Discussion

The BN feature mapping approach trained on IHM data outperforms the baseline IHM ASR system when recognizing the SDM test, whilst minor degradation is observed when recognizing the IHM test. When trained on SDM data, the mapping approach outperforms the baseline SDM ASR system when recognizing the IHM and SDM tests. Results reveal that DNN, employed to learn the feature mapping between the SDM and IHM conditions, improves distant speech recognition on the AMI meeting corpus. In addition, improved ASR performance on the RT-07 evaluation task indicates that the BN-based ASR systems, using acoustic models trained on SDM data, are robust to noise and reverberation. Furthermore, results show additional improvement when fMLLR is applied in combination with the mapping approach. Note that in [55], a denoising autoencoder is used for mapping, and ASR improvement is reported when recognizing noisy speech. One of the reasons for this is that they initialize the deep neural network by performing pre-training using an efficient algorithm [55]. In our preliminary experiments, direct feature mapping (i.e, SDM MFCC to IHM MFCC) does not yield any improvement when recognizing distant speech if we did not initialize the neural network.

As shown in Table 1, channel concatenation used in combination with the mapping approach for four channels yields slight performance degradation when compared to conventional beamforming. The performance is degraded by 2.1% (from 52.7% to 54.8%) absolute WER for 8-channel concatenation. However, these results are better than feature mapping using SDM, suggesting that the feature mapping approach can generalize to unseen conditions (i.e., the existing feature mapping network is used to forward-pass other channels). Also, eight channels improve over four channels by 0.5% absolute WER. An additional experiment is conducted by recognizing a single channel from the secondary array by using the existing feature mapping network (i.e., where a single channel of the primary array is used for mapping to an IHM condition). The result shows improvement by about 11.6% absolute WER (from 83.4% to 71.8%) when recognizing the SDM using the IHM ASR system. For comparison, the reported improvement for a single channel of the primary array is 12.6% absolute WER (from 76.0% to 63.4%). A better performance for channel concatenation using two and four microphones over the conventional beamforming is reported in [21]. Their experiments used tandem systems for concatenating features. From our experiments, the best strategy for feature mapping involving MDM

is to perform mapping using features from beamformed output. The improvements by 1.1% and 0.9% absolute WER (2% and 1.7% relative) are achieved by using four and eight microphones over conventional beamforming systems, respectively.

The best system is achieved by using the combination of 8-channel beamforming, feature mapping, and fMLLR. This system yields improvement by 8.1% (from 58.0% to 49.9%) absolute WER (14% relative) over the conventional SDM system. The same system yields improvement by 4.3% absolute WER (8% relative) over the SDM system with fMLLR. When a 4-gram LM is used, the best system obtains 47.4% WER (a further gain of 5% WER relative). This shows that the gains achieved by acoustic modeling are preserved with a stronger language model.

In regard to the multi-condition BN-based systems, the best performance for recognizing the IHM test resulted in a WER of 26.9% using the IHM model, and for recognizing the SDM test resulted in a WER of 52.2%, using the SDM model when fMLLR is applied. The best performance when recognizing IHM reported in this paper shows that multi-condition BN-based networks are inherently robust to noise. A small performance gain of 0.3% absolute WER is observed when recognizing SDM, as compared to the feature mapping approach (52.5% compared to 52.2%). The gains in performance from single channel case imply that the BN-based systems from multi-condition networks can be extended to the multi-channel case. Analysis of results from multi-microphones experiments will be left for future work.

7. Conclusions

This paper investigates DNN-based BN feature mapping using far-field microphones for improving distant-talking speech recognition on the AMI meeting corpus. For recognizing a mismatch condition, large improvement is observed when an acoustic model trained on IHM is used to recognize SDM data. The mapping approach is beneficial for improving distant speech recognition performance where the SDM acoustic model gives the best result for recognizing an SDM test, with a performance gain of 1.7% absolute WER. The ASR improvement obtained on the RT-07 evaluation task shows that the feature mapping approach generalizes to unseen conditions with a performance gain of 2.3% absolute WER. In terms of WER, the feature mapping approach is shown to be complementary to fMLLR feature space adaptation.

Feature mapping used in combination with beamforming and fMLLR improves the ASR performance over the baseline beamforming systems, while no improvement is observed when microphone channels are combined using feature concatenation. The multi-channel integration in DNN acoustic modeling may be beneficial for a small number of microphones since four-channel feature concatenation obtains roughly similar performance with beamforming. The ASR performance below 50% WER on AMI (SDM) evaluation set can be achieved by using the feature mapping approach in combination with beamforming and fMLLR. Experimental results on multi-style DNN training reveal that models trained using BN features extracted from a multi-condition network are inherently robust to noise. The BN-based system trained using the SDM data from such a network could be used to improve the performance of distant speech recognition.

8. Acknowledgment

This work was primarily supported by the European Community under the Eurostars project “DBox: A generic dialog box for multi-lingual conversational applications”. This work was also partially supported by the EC FP7 funding, under “Speaker Identification Integrated Project (SIIP)” and by the EC H2020 SESAR funding, under “Machine Learning of Speech Recognition Models for Controller Assistance (Malorca)” project. The authors would like to thank Dr. Blaise Potard and also the two anonymous reviewers and the Associate Editor for their valuable comments and suggestions.

References

[1] T. Yoshioka et al., Making Machines Understand Us in Reverberant Rooms: Robustness Against Reverberation for Automatic Speech Recognition, IEEE Signal Processing Magazine 29 (6) (2012) 114–126.

[2] B. D. Van Veen and K. M. Buckley, Beamforming: a versatile approach to spatial filtering, IEEE ASSP Magazine 5 (2) (1988) 2–24.

[3] X. Anguera, C. Wooters, and J. Hernando, Acoustic Beamforming for Speaker Diarization of Meetings, IEEE Transactions on Audio, Speech, and Language Processing 15 (7) (2007) 2011–2022.

[4] M. L. Seltzer, B. Raj, and R. M. Stern, Likelihood-maximizing beamforming for robust hands-free speech recognition, IEEE Transactions on Speech and Audio Processing 12 (5) (2004) 489–498.

[5] M. L. Seltzer and R. M. Stern, Subband Likelihood-Maximizing Beamforming for Speech Recognition in Reverberant Environments, IEEE Transactions on Audio, Speech, and Language Processing 14 (6) (2006) 2109–2121.

[6] C. Fox and T. Hain, Extending Limabeam with discrimination and coarse gradients, in: Proceedings of Interspeech, 2014.

[7] F. Metze et al., The ISL RT-04S meeting transcription system, in: Proceedings of the ICASSP-2004 Meeting Recognition Workshop, 2014.

[8] M. Wölfel and J. McDonough, Combining Multi-Source Far Distance Speech Recognition Strategies: Beamforming, Blind Channel and Confusion Network Combination, in: Proceedings of Interspeech, 2005.

[9] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, Speech recognition based on space diversity using distributed multi-microphone, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 1747–1750, 2000.

[10] M. Wolf and C. Nadeu, Channel selection measures for multi-microphone speech recognition, Speech Communication 57 (2014) 170–180.

[11] M. Harper, The Automatic Speech Recognition in Reverberant Environments (ASpIRE) Challenge, in: IEEE Automatic Speech Recognition and Understanding Workshop, 2015.

[12] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, The PASCAL CHiME Speech Separation and Recognition Challenge, Computer Science and Language 27 (2013) 621–633.

[13] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, The third CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines, in: IEEE Automatic Speech Recognition and Understanding Workshop, 2015.

[14] T. Hori et al., The MERL/SRI system for the 3rd CHiME challenge using beamforming, robust feature extraction, and advanced speech recognition, in: IEEE Automatic Speech Recognition and Understanding Workshop, 2015.

[15] S. Sivasankaran et al., Robust ASR using neural network based speech enhancement and feature simulation, in: IEEE Automatic Speech Recognition and Understanding Workshop, 2015.

[16] D. Yu and M. L. Seltzer, Improved Bottleneck Features Using Pretrained Deep Neural Networks, in: Proceedings of Interspeech, 2011.

[17] M. L. Seltzer, D. Yu, and Y. Wang, An investigation of deep neural networks for noise robust speech recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[18] F. Grézl, M. Karafiát, S. Kontár, and J. Černocký, Probabilistic and Bottleneck features for LVCSR of meetings, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2007.

[19] M. Karafiát et al., BUT BABEL system for spontaneous Cantonese, in: Proceedings of Interspeech, 2013.

[20] Y. Zhang, E. Chuangsuwanich, and J. Glass, Extracting Deep Neural Network Bottleneck Features using Low-Rank Matrix Factorization, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2014.

[21] Y. Liu, P. Zhang, and T. Hain, Using neural network front-ends on far field multiple microphones based speech recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2014.

[22] I. Himawan, P. Motlicek, D. Imseng, B. Potard, N. Kim, and J. Lee, Learning Feature Mapping using Deep Neural Network Bottleneck Features for Distant Large Vocabulary Speech Recognition, in: IEEE International Conference on Acoustic, Speech, and Signal Processing, 2015.

[23] T. Yoshioka, A. Ragni, and M. J. F. Gales, Investigation of Unsupervised Adaptation of DNN Acoustic Models with Filter Bank Input, in: Proceedings of Interspeech, 2014.

[24] P. Bell, P. Swietojanski, and S. Renals, Multi-level adaptive networks in tandem and hybrid ASR systems, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[25] P. Swietojanski and S. Renals, Learning Hidden Unit Contributions for Unsupervised Speaker Adaptation of Neural Network Acoustic Models, in: IEEE Spoken Language Technology Workshop, 2014.

[26] G. Saon et al., Speaker adaptation of neural network acoustic models using i-vectors, in: IEEE Workshop on Automatic Speech Recognition and Understanding, 2013.

[27] A. Senior and I. Lopez-Moreno, Improving DNN Speaker Independence with I-vector Inputs, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2014.

[28] Y. Miao, H. Zhang, and F. Metze, Towards Speaker Adaptive Training of Deep Neural Network Acoustic Models, in: Proceedings of Interspeech, 2014.

[29] Y. Miao, L. Jiang, H. Zhang, and F. Metze, Improvements to Speaker Adaptive Training of Deep Neural Networks, in: IEEE Spoken Language Technology Workshop, 2014.

[30] M. Miyoshi and Y. Kaneda, Inverse filtering of room acoustics, IEEE Transactions on Acoustics Speech and Signal Processing 36 (1988) 145–152.

[31] T. Nakatani, M. Miyoshi, and K. Kinoshita, Single-microphone blind dereverberation, chap. Speech enhancement, Springer Berlin Heidelberg, 247–270, 2005.

[32] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Transactions on Acoustics, Speech, and Signal Processing 27 (1979) 113–120.

[33] Y. Ephraim and D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing 32 (1984) 1109–1121.

[34] Y. Ephraim and D. Malah, Speech enhancement using minimum mean square log spectral amplitude estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing 33 (1985) 443–445.

[35] J. Meyer and K. U. Simmer, Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997.

[36] I. McCowan and H. Bourlard, Microphone array post-filter based on noise field coherence, IEEE Transactions on Speech and Audio Processing 11 (2003) 709–716.

[37] S. Makino, H. Sawada, and T. W. Lee, Blind Speech Separation, Springer Netherlands, 2007.

[38] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, A Regression Approach to Speech Enhancement Based on Deep Neural Networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (2015) 7–19.

[39] S. Srinivasan, N. Roman, and D. L. Wang, Binary and ratio time-frequency masks for robust speech recognition, Speech Communication 48 (2006) 1486–1501.

[40] A. Narayanan and D. L. Wang, Ideal Ratio Mask Estimation using Deep Neural Networks for Robust Speech Recognition, in: IEEE Conference on Acoustics, Speech, and Signal Processing, 2013.

[41] F. Weninger et al., Latent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings, chap. Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR, Lecture Notes in Computer Science, Springer International Publishing, 91–99, 2015.

[42] C. H. Knapp and G. C. Carter, The generalized correlation method for estimation of time delay, IEEE Transactions on Acoustics, Speech and Signal Processing 24 (4) (1976) 320–327.

[43] L. J. Griffiths and C. W. Jim, An alternative approach to linearly constrained adaptive beamforming, IEEE Transactions on An-

tennas and Propagation 30 (1) (1982) 27–34.
[44] J. Bitzer and K. U. Simmer, Microphone Arrays: Signal Processing Techniques and Applications, chap. Superdirective Microphone Arrays, Springer Berlin Heidelberg, 19–38, 2001.
[45] A. Stolcke et al., The SRI-ICSI Spring 2007 Meeting and Lecture Recognition System, Lecture Notes in Computer Science 4625 (2008) 450–463.
[46] T. Hain et al., Transcribing Meetings With the AMIDA Systems, IEEE Transactions on Audio, Speech and Language Processing 20 (2) (2012) 486–498.
[47] P. Swietojanski, A. Ghoshal, and S. Renals, Hybrid acoustic models for distant and multichannel large vocabulary speech recognition, in: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2013.
[48] D. Marino and T. Hain, An Analysis of Automatic Speech Recognition with Multiple Microphones, in: Proceedings of Interspeech, 2011.
[49] S. Kim and I. Lane, Recurrent Models for Auditory Attention in Multi-Microphone Distance Speech Recognition, in: arXiv preprint arXiv:1511.06407, 2015.
[50] P. Swietojanski, A. Ghoshal, and S. Renals, Convolutional Neural Networks for Distant Speech Recognition, IEEE Signal Processing Letters 9 (2014) 1120–1124.
[51] Y. Zhang et al., Highway Long Short-Term Memory RNNs for Distant Speech Recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2016.
[52] B. de Vries et al., Neural Network Speech Enhancement for Noise Robust Speech Recognition, in: International Workshop on Applications of Neural Networks to Telecommunications 2, 1995.
[53] Q. Lin, C. Che, D.-S. Yuk, L. Jin, B. de Vries, J. Pearson, and J. Flanagan, Robust distant-talking speech recognition, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996.
[54] W. Li, J. Dines, M. Magimai-Doss, and H. Bourlard, Non-linear mapping for multi-channel speech separation and robust overlapping spech recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009.
[55] X. Feng, Y. Zhang, and J. Glass, Speech Feature Denoising and Dereverberation via Deep Autoencoders for Noisy Reverberant Speech Recognition, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2014.
[56] W. Li, L. Wang, Y. Zhou, J. Dines, M. Magimai-Doss, H. Bourlard, and Q. Liao, Feature Mapping of Multiple Beamformed Sources for Robust Overlapping Speech Recognition Using a Microphone Array, in: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014.
[57] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, Single-Channel Mixed Speech Recognition using Deep Neural Network, in: IEEE International Conference on Acoustic, Speech and Signal Processing, 2014.
[58] A. Ghoshal and P. Swietojanski, and S. Renals, Multilingual training of deep neural networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.
[59] D. Povey et al., The Kaldi speech recognition toolkit, in: Automatic Speech Recognition and Understanding, 2011.
[60] P. Motlicek, D. Povey, and M. Karafiát, Feature and Score Level Combination of Subspace Gaussians in LVCSR Task, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.
[61] Y. M. Cheng et al., A Robust Front-End Algorithm for Distributed Speech Recognition, in: Proceedings of Eurospeech, 2001.